一种基于 KNN 算法的客户身份识别方法

杨 菁,刘鲲鹏,金 鹏
(国网客服中心服务考评部,天津 300309)

摘 要:针对电力客户标签对客户的认知不清晰,客服业务大多针对自然人开展,客户标签标记在电话号码上,而传统电力业务主要针对户(户号)开展,客户标签标记在户号上,存在信息无法共享的困难,提出了基于 95598 业务,利用大数据分析及文本挖掘方法,构建统一身份识别模型,有效识别客户来电号码与户号的对应关系。采用分词技术,有效解析用电地址信息、客户姓名等内容,并计算地址相似度得分、姓名相似度得分,作为对应关系校验以及识别疑似户号的关键因子指标;针对能获取到的对应关系,构建权重划分模型,计算对应关系匹配度得分,根据分值大小,校验对应关系的可靠性;针对找不到户号对应关系的来电号码,基于文本相似度得分构建 KNN 模型,计算对应关系匹配度得分,依据分值大小,识别疑似户号。

关键词:统一身份识别;文本挖掘;大数据;KNN 模型
文章编号:2096 - 4633(2019)04 - 0067 - 07 **中图分类号**:TM744 **文献标志码**:B

95598 客服人员在受理客户来电业务时,需要对客户户号进行核对,以便解决客户来电诉求。由于大部分客户无法提供客户户号,坐席人员需要通过询问客户用电地址信息与现有档案用电地址进行匹配,获取客户户号。目前客服中心标签[1-7]是以电话号码为对象构建,省公司标签是以用户号为对象构建,为实现中心和省公司标签共享,需要构建电话号码和用户号之间的动态精准匹配关系,支撑以电话号码为对象的客户画像和以用户号为对象的客户画像,实现中心和省公司在标签对象上的融合应用。鉴于此,识别客户来电号码与户号的对应关系势在必行。

基于 95598 数据,客户电话号码与户号的对应关系划分为以下两种情况:

①有号码有户号:针对这种情况,需要提取工单中号码和户号的对应关系,并与客户档案中的对应数据进行匹配、校验,识别号码与户号对应关系。

②有号码无户号:此类情况分为号码记录在档案,号码未记录在档案两种。

号码记录在档案:根据用电地址信息、客户姓名、供电单位、来电频次、最近来电时间等校验因素判别对应关系的准确度;

号码未记录在档案:根据用电地址信息、客户姓名、供电单位、来电频次、最近来电时间等因素,引入大数据模挖掘术,通过文本挖掘、构建模型,识别疑似户号。

针对上述对应关系,从多个分析维度出发进行数据探索,寻找号码与户号的对应关系,引入大数据分析挖掘技术,对每一类关系计算匹配度得分,综合每一类关系得分,针对一户多号、一号多户情况划分优先级。

1 识别方法

1.1 对应关系识别

1.1.1 有号码有户号

1.1.1.1 数据源

近两年浙江省内户号不为空的 95598 工单数据;客户档案数据。

1.1.1.2 研究步骤

(1)数据加工:提取 95598 工单业务中记录户号与号码的工单,并加工关系基表(户号、号码、来电频次、时间点、地址、姓名等);

(2)正确性校验:为保证对应关系的准确性,对提取的对应关系进行数据校验,排除无效关系。

校验准则:在进行多维度因素校验的同时,需计算对应关系的校验得分,即关系匹配度得分,根据分值确定关系可靠性。

①对应关系在档案中成功匹配,无需进行校验,规定此种情况匹配度得分最高;

②对应关系在档案没有匹配成功:此类情况需对工单数据中的用电地址信息、客户姓名与档案进行准确性校验。同时结合用户拨打行为特征,如拨打频次、最近拨打时间、来电时间点、拨打业务类型等行为综合

校验对应关系的准确性。在数据校验的过程中,计算关系匹配度得分,根据匹配度得分确定关系可靠性。

计算匹配度得分:需借助于大数据文本挖掘技术,对涉及到文本校验因素,进行分词并计算文本相似度,进而将文本相似度作为因素指标;对拨打行为指标(如号码拨打次数、拨打时间点、最近拨打时间、拨打事件类型、用电地址在历史工单中出现次数、客户姓名在历史工单中出现次数、同一户号是否在历史工单中出现,出现该户号的频次等因素)可作为行为量化因素指标;通过使用层次分析法[8-10]、熵值法[11-12]、因子分析法等大数据建模方法,构建指标权重划分模型,对其计算各个因子指标权重,进而计算关系匹配度得分。

说明:在校验过程中,若不满足以上因素的检验条件,则将不满足条件的来电号码放到下一分类情况(即有号码无户号情况)进行关系识别。

1.1.1.3 流程图展示

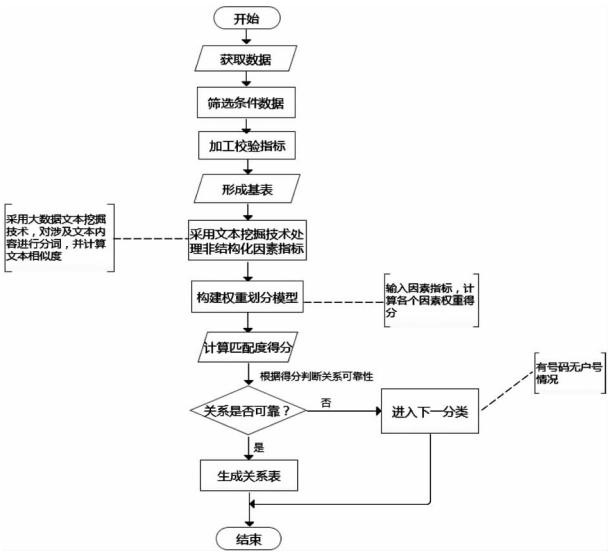

图 1 识别过程 1

Fig. 1 Recognition procedure 1

1.1.2 有号码无户号(号码记录在档案)

1.1.2.1 数据源

近两年户号为空且来电号码在档案中有记录的 95598 工单数据;

提取步骤 1.1.1 中判别无效关系且号码出现在档案的 95598 工单;

客户档案数据。

1.1.2.2 研究步骤

(1)数据加工:提取近两年户号为空且来电号码在档案中有记录的 95598 工单,并通过电话号码获取档案中的户号;

提取步骤 1.1.1 中判别无效关系且号码记录在档案的工单;

加工 95598 工单数据基表(记录工单编号、来电号码、客户编号、用电地址、客户姓名、拨打频次、最近拨打时间等内容)、客户档案数据基表(记录客户编号、用电地址、客户姓名、联系号码等内容);

(2)正确性校验:校验准则:此类情况需对工单数据中的用电地址信息、客户姓名与档案进行准确性校验。同时,再结合用户拨打行为特征,如拨打频次、最近拨打时间、来电时间点、拨打业务类型等行为综合校验对应关系的准确性。在数据校验的过程中,计算关系匹配度得分,根据匹配度得分确定关系可靠性。

计算匹配度得分:需借助于大数据文本挖掘技术,对涉及到文本校验因素,进行分词并计算文本相似度,进而将文本相似度作为因素指标;对拨打行为指标(如号码拨打次数、拨打时间点、最近拨打时间、拨打事件类型、用电地址在历史工单中出现次数、客户姓名在历史工单中出现次数、同一户号是否在历史工单中出现,出现该户号的频次等因素)可作为行为量化因素指标;通过使用层次分析法、熵值法、因子分析法等大数据建模方法,构建指标权重划分模型,对其计算各个因子指标权重,进而计算关系匹配度得分。

在校验过程中,若不满足以上因素的检验条件,则将不满足条件的来电号码放到下一分类情况(即有号码无户号且号码未记录在档案情况)进行关系识别。

1.1.2.3 流程图展示

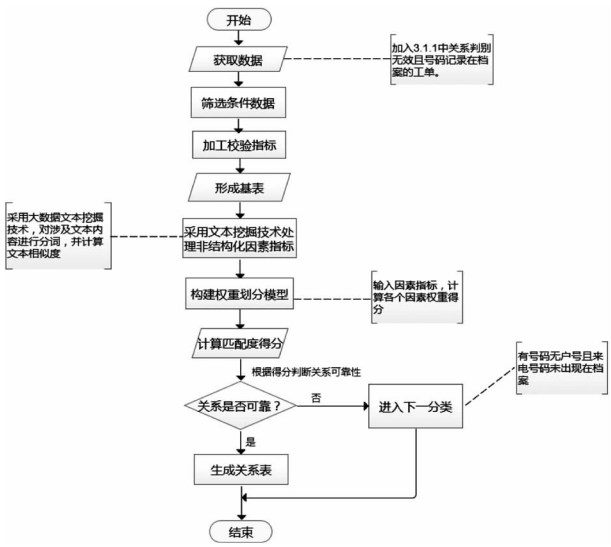

图 2 识别过程 2

Fig. 2 Recognition procedure 2

1.1.3　有号码无户号(号码未记录在档案)

此类情况由于来电号码未记录在档案中,无法通过电话号码获取相应的户号,因此需要引入文本挖掘,通过对客户工单中的地址信息与档案中的地址信息进行分析,构建文本相似模型,识别疑似户号。

1.1.3.1　数据源

近两年户号为空且来电号码在档案中没有记录的 95598 工单数据;

提取步骤:1.1.1 中判别无效关系且号码未记录在档案的 95598 工单;

提取步骤:1.1.2.1 中判别无效关系的号码工单;

客户档案数据。

1.1.3.2　研究步骤

(1)数据加工:提取近两年户号为空且来电号码在档案中没有记录的 95598 工单;

提取步骤 1.1.2.1 中判别无效关系的工单;

加工 95598 工单数据基表(工单编号、来电号码、客户编号等)、客户档案数据基表(客户编号、用电地址、客户姓名、联系号码等);

(2)因素指标:在寻找疑似户号的过程中,需要构建因子指标,判别待识别来电客户的通话行为、身份信息、地址信息等因素是否与现存对应关系的行为一致或者相近,最终寻找此来电号码的疑似户号。现存对应关系可分为两类:①基于步骤 3.1.1 和 3.1.2.1 识别出的对应关系;②其余的为档案数据中已存在的对应关系。基于以上数据源,因子指标按照数据结构分为非结构化指标与结构化指标两类。

非结构化指标:客户用电地址、客户姓名、受理内容中提取信息量(户号、电话号码、姓名等)、处理意见中提取的信息量(户号、电话号码、姓名等)等文本内容;

结构化指标:来电频次、来电时间点、通话时长、各个业务类型的来电频次、来电时长以及最近来电时间等通话行为。

(3)数据建模识别户号。

①非结构指标相似度计算:非结构化指标相似度计算方法:基于以上几类文本数据,采用大数据文本挖掘技术,对其进行文本分词,进而将非结构化数据转化为结构化处理。将分词之后的各个文本内容根据出现频次,构造向量空间,利用余弦夹角度量方法、最长公共子序列方法、最小边际距离算法等,计算各个对应文本的相似度,相似度作为建模因子指标。

②结构化指标构建 KNN 模型:通过输入非结构指标(即文本挖掘计算出的相似度),以及结构化因素指标,构建 KNN[11-15] 数据模型计算每个号码对象与现存对应关系的相似度,最终来确定该号码对应的疑似户号,实现号码与户号的匹配。现存对应关系可分为两类:①基于步骤 1.1.1 和 1.1.2.1 识别出的对应关系;②其余的为档案数据中已存在的对应关系。

在训练 KNN 模型的同时,需确定出合适的 K 值作为户号类别归属的判别,在筛户号归属的同时,需遵从如下原则:

在邻近的 K 个可选户号归属中,若属于 1.1.1 与 1.1.2.1 的对应关系优先选取该户号(号码关系相对可靠),否则按照模型相似度得分来分配疑似户号归属。

1.1.3.3　流程图展示

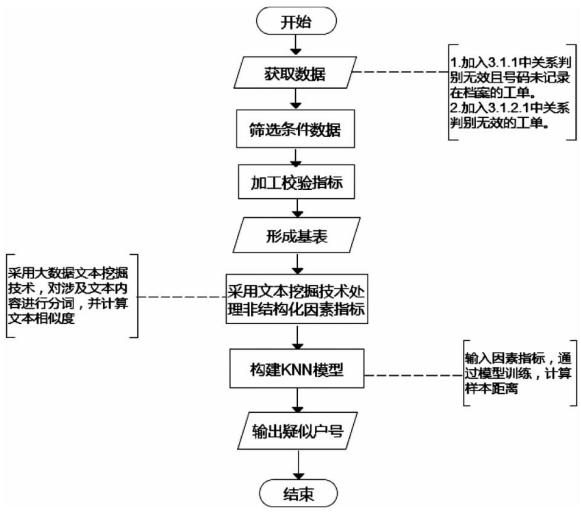

图 3　识别过程 3

Fig. 3　Recognition procedure 3

综述:综合 1.1.1 与 1.1.2 两类情况找寻的户号信息,进行合并处理,形成户号与号码的对应关系。在合并后的对应关系中,对应关系存在如下三种情况:

号码与户号 1 对 1;号码与户号 1 对多;号码与户号多对 1。

针对号码与户号多对多的情况,需进行优先级划分。

1.2 优先级划分

对于一户多号、一号多户的对应关系,需制定关系优先级,选取最可靠的对应关系。制定如下规则对其进行优先级划分:

(1)针对 1.1.1(有号码有户号)分类情况,按照匹配度得分,选取一户多号、一号多户最为可靠的关系;

(2)针对 1.1.2.1(有号码无户号且号码出现在档案)分类情况,按照匹配度得分,选取一户多号、一号多户最为可靠的关系;

(3)针对 1.1.2.2(有号码无户号且号码未出现在档案)分类情况,按照模型相似度得分,选取一户多号、一号多户最为可靠的关系;

综合三部分对应关系,针对合并之后出现一户多号、一号多户的情况再次进行优先级划分,划分规则遵从如下规定:

满足条件(1)的对应关系优先级最高;

满足条件(2)的对应关系优先级次之;

满足条件(3)的对应关系优先级最低;

2 模型因子指标设计

针对寻找到的对应关系需进行关系校验,通过文本挖掘算法计算文本相似度得分,进而将文本相似度得分以及拨打行为指标作为构建权重划分模型的输入因子,通过模型计算相似度得分,校验关系可靠性。

表 1 模型数据因子

Tab. 1 Model data

序号	宽表维度	宽表明细
1	文本相似度得分	地址相似度得分
2		姓名相似度得分
3	拨打原因	催办各业务主体平均拨打次数(近一个月、近三个月、近一年内)
4		投诉各业务主体平均拨打次数(近一个月、近三个月、近一年内)
5		催办各业务主体平均拨打时长(近一个月、近三个月、近一年内)
6		投诉各业务主体平均拨打时长(近一个月、近三个月、近一年内)
7	拨打偏好时段	22:00 – 次日 7:00 拨打次数(近一个月、近三个月、近一年内)
8		7:00 – 次日 12:00 拨打次数(近一个月、近三个月、近一年内)
9		12:00 – 次日 17:00 拨打次数(近一个月、近三个月、近一年内)
10		17:00 – 次日 22:00 拨打次数(近一个月、近三个月、近一年内)
11		周末拨打次数(近一个月、近三个月、近一年内)
12		元旦假日拨打次数(近一年内)
13		春节假日拨打次数(近一年内)
14		清明节假日拨打次数(近一年内)
15		劳动节假日拨打次数(近一年内)
16		端午节假日拨打次数(近一年内)
17		中秋节假日拨打次数(近一年内)
18		国庆假日拨打次数(近一年内)
19		2 – 3 月拨打次数(近一年内)
20		4 – 6 月拨打次数(近一年内)
21		7 – 8 月拨打次数(近一年内)
22		9 月拨打次数(近一年内)
23		10 月 – 次年 1 月拨打次数(近一年内)

序号	宽表维度	宽表明细
24	户号号码对应关系的指标	对应关系来电总次数
25		对应关系近 3 月来电次数
26		对应关系近 6 月来电次数
27		平均通话时长
28		最小通话时长
29		最大通话时长
30		拨打事件数
31		最近来电时间
32		历史拨打记录中是否存在该对应关系
33		来电时间点(工作日、非工作日、节假日)
34	户号的历史拨打行为数据	户号来电总次数
35		户号来电平均通话时长
36		户号来电最大通话时长
37		户号来电最小通话时长
38		最近记录号码是否为对应号码
39	号码的历史拨打行为数据	号码来电总次数
40		号码来电平均通话时长
41		号码来电最大通话时长
42		号码来电最小通话时长
43		最近记录户号是否为对应户号

3　模型效果评估

选取浙江省 2016/04/01 至 2017/04/01 工单数据作为建模数据,通过构建权重划分模型,计算对应关系匹配度得分。选取未来 5 个月内(2017/04/01 -2017/08/31)有过拨打且记录客户户号的工单作为模型验证集,对模型输出结果进行关系验证,并将数据作十分位,分别验证模型模型的命中率、覆盖率情况,验证结果如表 2 所示。

表 2　模型匹配结果

Tab. 2　Model matching results

得分排名	关系数量	累计关系数量	匹配关系数	累计匹配关系数	累计命中率/%	累计覆盖率/%
10	2 958	2 958	1 824	1 824	61.66	17.36
20	2 958	5 916	1 555	3 379	57.12	32.16
30	2 958	8 874	1 266	4 645	52.34	44.21
40	2 958	11 832	1 156	5 801	49.03	55.22
50	2 958	14 790	1 069	6 870	46.45	65.39
60	2 958	17 748	983	7 853	44.25	74.75
70	2 958	20 706	851	8 704	42.04	82.85
80	2 958	23 664	657	9 361	39.56	89.10
90	2 958	26 622	624	9 985	37.51	95.04
100	2 958	29 580	521	10 506	35.52	100.00

从上表可知,分值越高,模型命中率越高,符合分值越高,户号与号码对应关系越紧密的趋势特征;模型命中率在分值排名前 70% 以内都高于 40%,并且在分值前 70% 模型覆盖率达到了 82.5%,故建议选取分值排名前 70% 作为模型目标关系群。

4 结论

从浙江省近 1 年受理工单情况入手,通过构建统一身份识别模型,共有效识别出对应关系 272 万,涉及工单量为 431.74 万工单,覆盖浙江省近 1 年工单总量的 67.35%,即有 67.35% 的受理工单通过模型有效识别出户号。

参考文献:

[1] 厉建宾,朱雅魁,吴彬彬,等. 电力客户标签体系框架构建研究及应用实践[J]. 计算机工程与应用,2017,53(S1):229 - 235.
LI Jianbin, ZHU Yakui, WU Binbin, et al. Research and application of power customer tag system framework[J]. Computer Engineering and Application,2017,53(S1):229 - 235.

[2] 林森,欧阳柳. 基于大数据理论的电力客户标签体系构建[J]. 电气技术,2016,(12):98 - 101,112.
LIN Sen, OUYANG Liu. Study on the construction of power customer label system based on big data theory[J]. Electrical Engineering,2016,(12):98 - 101,112.

[3] 史梦洁,王庆娟,涂莹,等. 电力营销客户标签体系分类方法研究[J]. 电力需求侧管理. 2018,20(02):51 - 53.
SHI Mengjie, WANG Qingjuan, TU Ying, et al. Research on classification method of power customer tag collection[J]. Power Demand Side Management,2018,20(02):51 - 53.

[4] 裘华东,涂莹,丁麒. 基于标签库系统的电力企业客户画像构建与信用评估及电费风险防控应用[J]. 电信科学,2017,(Z1):206 - 213.
QIU Huadong,TU Ying,DING Qi. Construction of power customer portrait and its credit evaluation and electricity fee risk control based on tag library system[J]. Telecommunications Science,2017,(Z1):206 - 213.

[5] 杜伟,蒋鹏,王文浩,等. 基于变压器大数据画像技术与应用研究[J]. 电力大数据,2017,20(08):10 - 14.
DU Wei, JIANG Peng, WANG Wenhao, et al. Research on technology and application of big data portrait based on transformer[J]. Power Systems and Big Data,2017,20(08):10 - 14.

[6] 吕辉,许道强,仲春林,等. 基于电力大数据的标签画像技术与应用研究[J]. 电力信息与通信技术,2017,15(02):43 - 48.
LV Hui, XU Daoqiang, ZHONG Chunlin, et al. Study on tag portrait technology based on electric power big data and its application[J]. Electric Power Information and Communication Technology,2017,15(02):43 - 48.

[7] 谢骏凯,徐千,丁炳淼. "客户画像"在电费回收风险管控中的应用[J]. 电力需求侧管理. 2016. 18(S1):74 - 76.
XIE Jukai, XU Qian, DING Bingmiao. Application of customer models in the risk control of electricity fees' recovery[J]. Power Demand Side Management,2016,18(S1):74 - 76.

[8] 邓雪,李家铭,曾浩健,等. 层次分析法权重计算方法分析及其应用研究[J]. 数学的实践与认识,2012,42(07):93 - 100.
DENG Xue, LI Jiaming, ZENG Haojian, et al. Research on computation methods of AHP Wight Vector and its applications[J]. Mathematics in practice and theory,2012,42(07):93 - 100.

[9] 金菊良,魏一鸣,丁晶. 基于改进层次分析法的模糊综合评价模型[J]. 水利学报,2004,(03):65 - 70.
JIN Juliang, WEI Yiming, DING Jing. Fuzzy comprehensive evaluation model based on improved analytic hierarchy process[J]. Journal of Hydraulic Engineering,2004,(03):65 - 70.

[10] 吴殿廷,李东方. 层次分析法的不足及其改进的途径[J]. 北京师范大学学报(自然科学版),2004,40(02):264 - 268.
WU Dianting, LI Dongfang. Shortcomings of analytical hierarchy process and the path to improve the method[J]. Journal of Beijing Normal University(Natural Science),2004,40(02):264 - 268.

[11] 郭显光. 改进的熵值法及其在经济效益评价中的应用[J]. 系统工程理论与实践,1998,18(12):99 - 103.
GUO Xianguang. Application of improved entropy method in evaluation of economic result[J]. Systems Engineering-Theory & Practice,1998,18(12):99 - 103.

[12] 孙刘平,钱吴永. 基于主成分分析法的综合评价方法的改进[J]. 数学的实践与认识,2009,39(18):15 - 20.
SUN Liuping, QIAN Wuyong. An improved method based on principal component analysis for the comprehensive evaluation[J]. Practice and Understanding of Mathematics,2009,39(18):15 - 20.

[13] 代六玲,黄河燕,陈肇雄. 中文文本分类中特征抽取方法的比较研究[J]. 中文信息学报,2004,18(01):26 - 32.
DAI Liuling, HUANG Heyan, CHEN Zhaoxiong. A comparative study on feature selection in chinese text categorization[J]. Journal of Chinese Information Processing,2004,18(01):26 - 32.

[14] 李蓉,孙媛. SVM-KNN 分类器——一种提高 SVM 分类精度的新方法[J]. 科学技术与工程,2009,9(16):4653 - 4656.
LI Rong,SUN Yuan. Application of SVM-KNN classifier into web page classification[J]. Science Technology and Engineering,2009,9(16):4653 - 4656.

[15] 罗辛,欧阳元新,熊璋,等. 通过相似度支持度优化基于 K 近邻的协同过滤算法[J]. 计算机学报,2010,33(08):1437 - 1445.
LUO Xin,OUYANG Yuanxin,XIONG Zhang,et al. The effect of similarity support in K-nearest-neighborhood based collaborative filtering[J]. Chinese Journal of Computers,2010,33(08):1437

－1445.

[16] 刘义,景宁,陈荦,等. MapReduce 框架下基于 R－树的 k－近邻连接算法[J]. 软件学报,2013,24(08):1836－1851.
LIU Yi, JING Ning, CHEN Qian, et al. Algorithm for processing K-Nearest join based on R-Tree in MapReduce[J]. Journal of Software,2013,24(08):1836－1851.

[17] 王振军,黄瑞章. 基于 Spark 的矩阵分解与最近邻融合的推荐算法[J]. 计算机系统应用,2017,26(04):124－129.
WANG Zhenjun, HUANG Ruizhang. Recommendation algorithm using matrix decomposition and nearest neighbor fusion based on Spark[J]. Computer Systems & Applications,2017,26(04):124－129.

收稿日期:2018－12－03

作者简介:

杨　菁(1989),女,理学博士,工程师,主要从事电力客户标签研究应用相关工作。

(本文责任编辑:施　玉)

A customer identification method based on KNN algorithm

YANG Jing, LIU Kunpeng, JIN Peng

(State Grid Service Evaluation Department of Customer Service Center, Tianjin 300309 China)

Abstract: In order to solve the problems such as lack of clear understanding of customers, customer service business is mostly carried out for natural persons, customer labels are marked on telephone numbers, while traditional electric power business is mainly carried out for households (household numbers). Customer labels are marked on household numbers, which makes it difficult to share information. Based on the 95598 business, a unified identity recognition model is constructed by using big data analysis and text mining to effectively identify the corresponding relationship between customer phone number and account number. Using participle technique, efficient parsing address information, customer name, address and calculate the similarity score, name similarity scores, check the corresponding relationships and identify the key factor of the suspected door number indicator. According to the obtained correspondence, the weight division model is built, the matching score of corresponding relation is calculated, and the reliability of corresponding relation is verified according to the score value. KNN model was built based on the similarity score of text to calculate the matching score of corresponding relation, and identify the suspected account number according to the score value.

Key words: unified identity recognition; text mining; big data; KNN model